One way analysis of Variance

Yesoda Bhargava, MPH, University of York, UK.

February 14, 2020

Introduction

- Many of the most effective statistical studies are comparative. Comparisons could be shown by side-by-side boxplots.
- Now the question we ask is, “Is the difference between groups statistically significant? ”
- When only two groups are compared we can use two-sample t-tests procedures. Now, we wish to compare any number of means by techniques that generalize the two-sample t-test and share its robustness and usefulness.
- The statistical methodology for comparing several means is called analysis of variance, ir simply ANOVA.
- When there is only one way to classify the population of interest, we use one-way ANOVA to analyse the data. For eg. compare the mean birth weight by different ethnic groups.
- In many situations there are more than one way to classify the populations.

Dataset used

- For this lecture, we use the Melanoma dataset in R. To load the data:
  library(MASS)
  data=birthwt

One-way ANOVA

- If we have random samples from the two populations, we ocmpute a two-sample t-statistic and its P-value to assess the statistical significance of the difference in the sample means.
- Comparison of several means is done in the same way. Instead of a t-statistic, ANOVA uses F Statistic and its P-value to evaluat the null hypothesis that all of several population means are equal.
- The question we ask in ANOVA is, "Do all groups have the same **population** mean? "To answer this question we compare the sample means.
- The purpose of ANOVA is to assess whether the observed differences among sample means are statistically significant. Could a variation this large be plausibly due to chance, or is it good evidence among for a difference among the population means? This question cannot be answered from the sample means alone.

- Because the standard deviation of a sample mean $\bar{x}$ is the population standard deviation σ divide by $\sqrt{n}$, the answer depends upon both the variation within the groups of observations and the sizes of the samples.
- To begin with we look at the boxplots, they are a good preliminary display of ANOVA data.

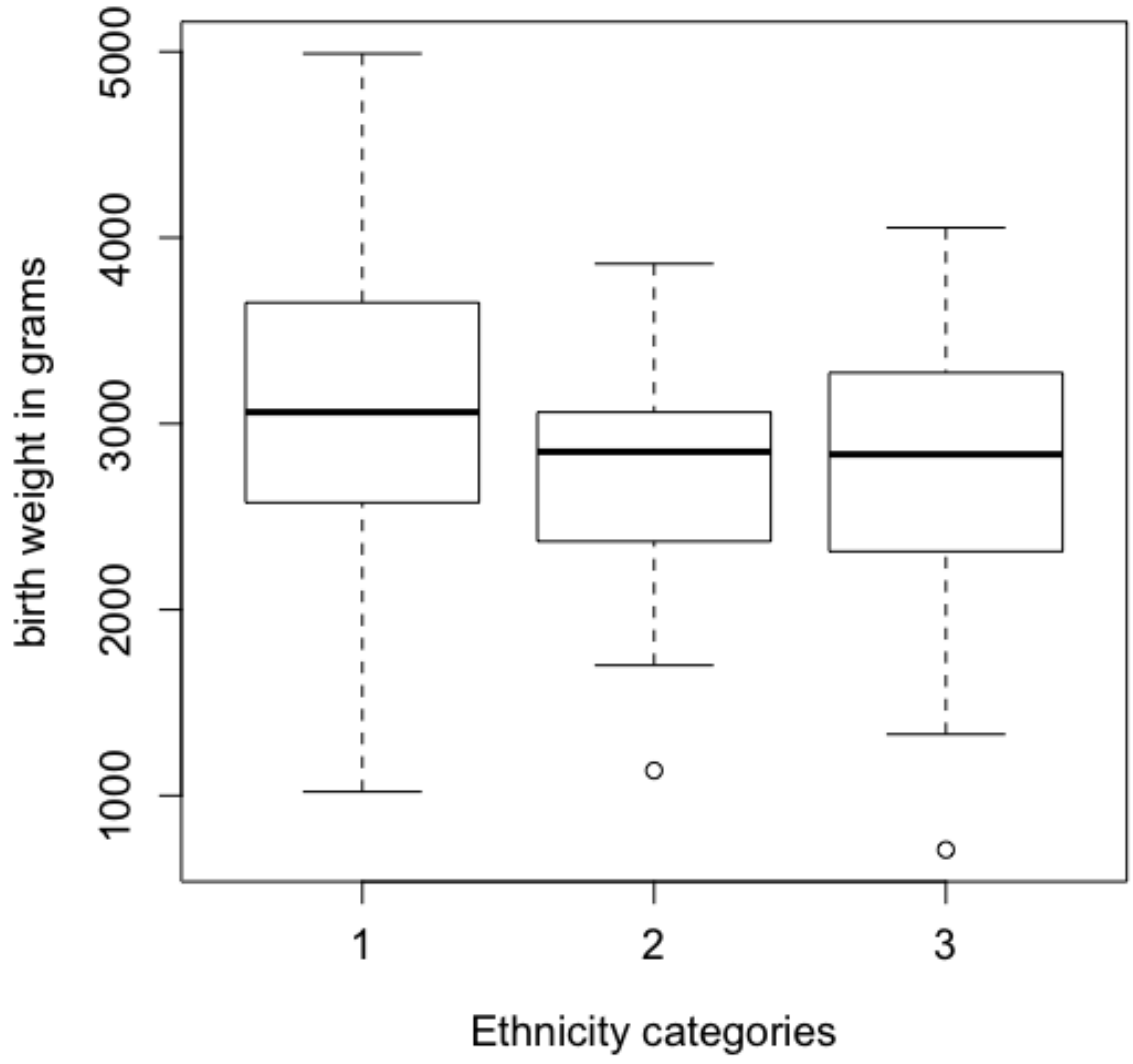


Figure 1: Side-by-side boxplot for birth weight versus ethnicity.

The two-sample t statistic

Two-sample t-statistics compare the means of two populations. If the two populations are assumed to have equal but unknown standard deciations and the sample sizes are both equal to n, the t-statistic is:

$$t = \frac{\bar{x} - \bar{y}}{s_p \times \sqrt{\frac{1}{n} + \frac{1}{n}}} \tag{1}$$

The square of this t-statistic is:
$t^2 = \frac{\frac{n}{2} \times (\bar{x} - \bar{y})^2}{s_p^2}$

- If we use ANOVA to compare two populations, the ANOVA F statistic is exactly equal to this t^2.
- The numerator in the t^2 statistic measures the variation between the groups in terms of the difference between their sample means $\bar{x}$ and $\bar{y}$.
- It includes a factor for the common size n. The numerator can be large because of a large difference between the sample means or because the sample sizes are large.

- The denominator measures the variation within groups by s_p^2, the pooled estimator of the common variance.
- If the within-group variance is small, the same variation between the groups produces a larger statistic and a more significant result.
- To assess whether several populations all have the same mean, we compare the variation among the means of several groups with the variation within groups. Because we are comparing variation, the method is called analysis of variance.

An overview of ANOVA

- ANOVA tests the null hypothesis that the population means are equal. The alternative is that they are not all equal.
- This alternative could be true because all of the means are different or simply because one of them differs from the rest.
- If we reject the null hypothesis, we need to perform some further analysis to draw conclusions about which population means differ from which others.

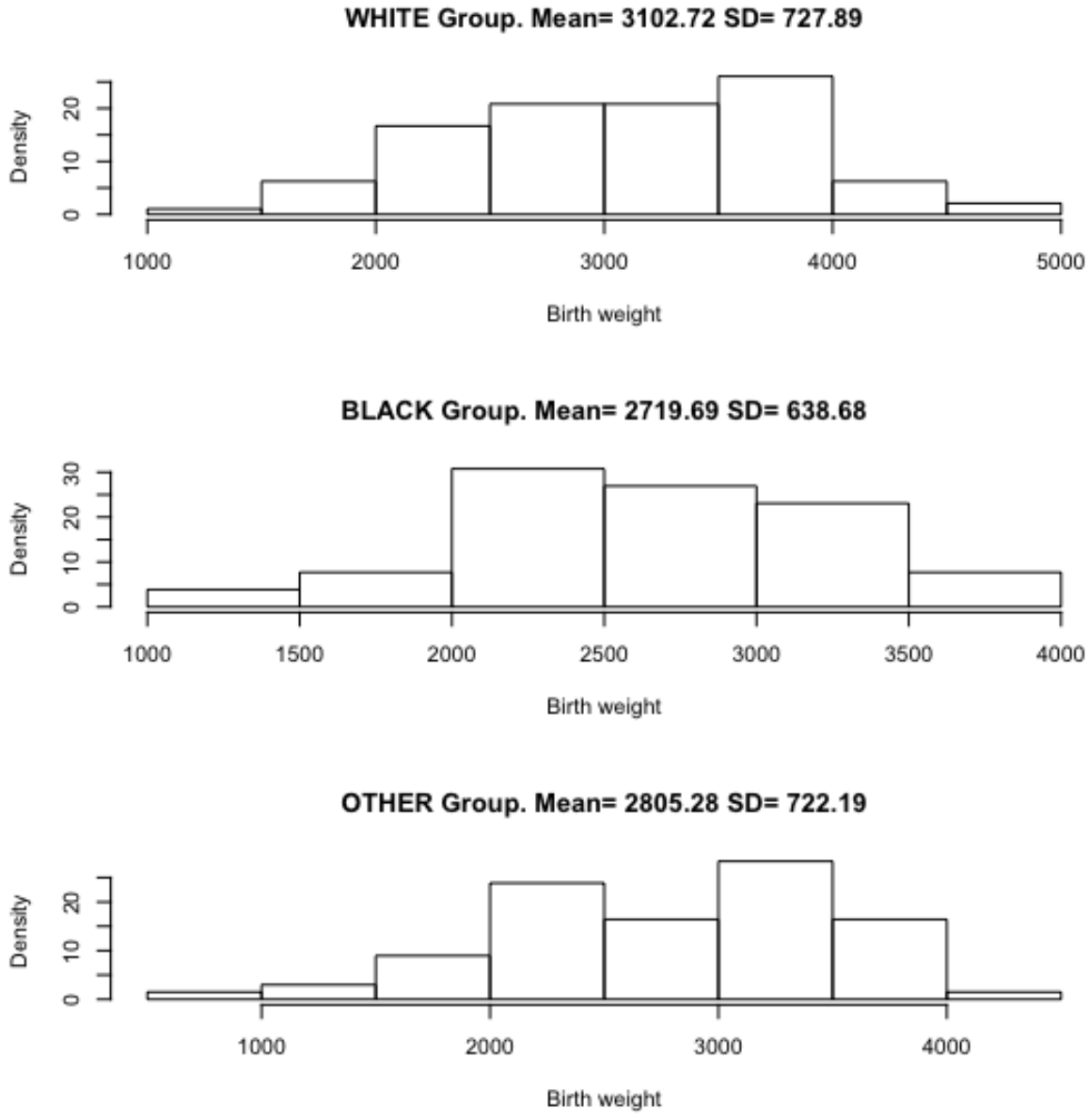


Figure 2: Histograms and descriptive statistics for the birth-weight example.

Table 1: Descriptive statistics for the birth weight example

Ethnicity category	n	$\bar{x}$	s
White	96	3102	727.89
Black	26	2719	638
Other	67	2805	722

Sample size for Black category is small. To assume normal distribution of variables look at the QQ plot.

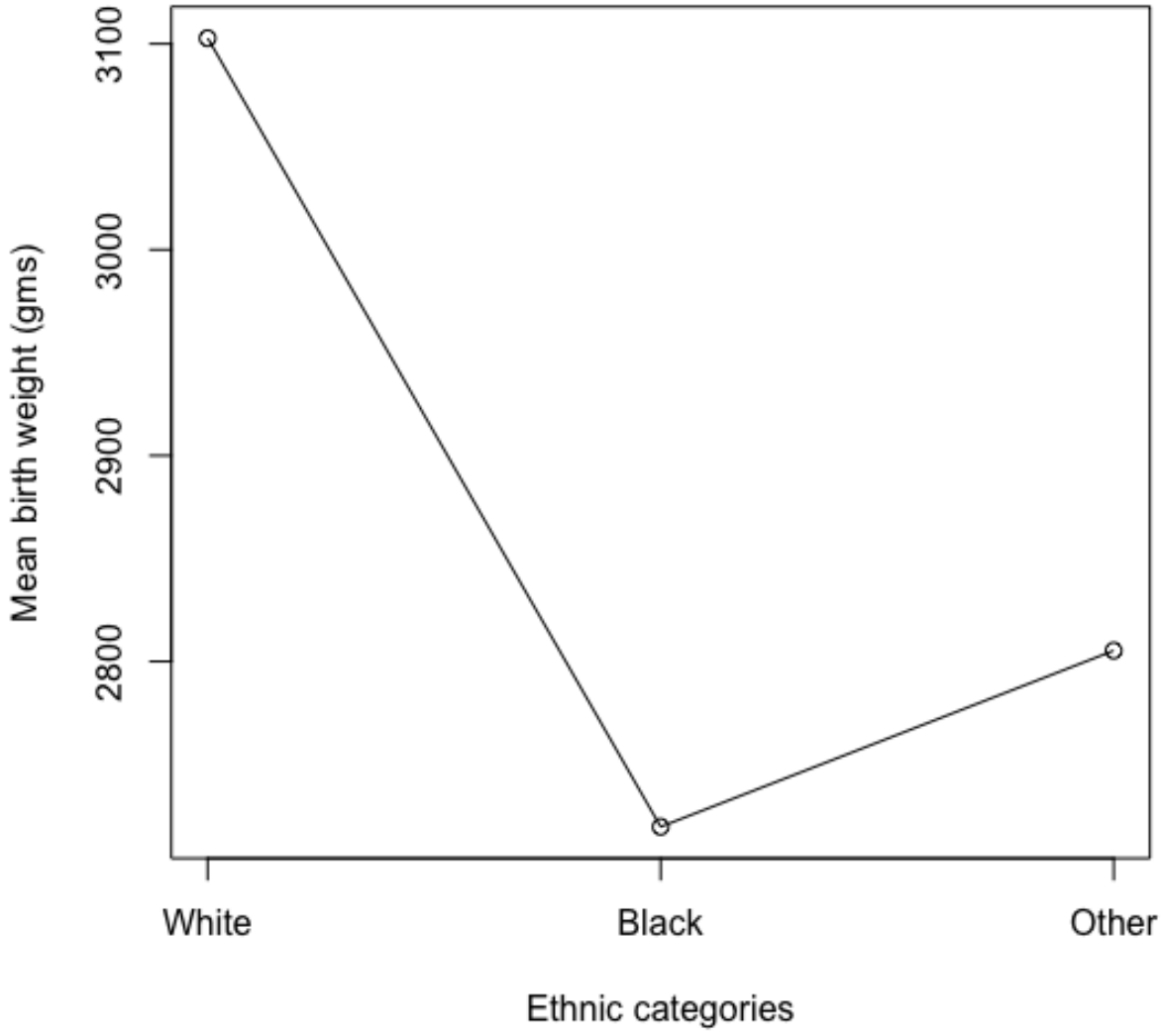


Figure 3: Birth weight means for the birth weight example.

- It appears that the Black and other ethnic category are similar, while the White babies have a slightly different mean.
- To apply ANOVA in this setting, we view the three samples that we have as three independent random samples from three distinct populations. Each of these populations has a mean and our inference asks questions about these means.
- We first ask whether or not there is sufficient evidence in the data to conclude that the corresponding population means are not all equal.
- H_0 : Population mean birth weight is the same for all three groups of ethnicities. The alternative is that they are not all the same.
- Rejecting the null hypothesis that the means are not all the same using ANOVA is not the same as concluding that all of the means are different from one another. Additional analysis is required to distinguish among these possibilities.

- When there are particular versions of the alternative hypothesis that are of interest, we use contrasts to examine them. In our example, we might want to compare the White ethnicity babies with all of the other babies.
- Note that to use contrasts, it is necessary that the questions of interest be formulated before examining the data. It is cheating to make up these questions after analyzing the data.
- If we have no specific relations among the means in mind before looking at the data, we instead use a multiple comparisons procedure to determine which pairs of population means differ significantly.

The ANOVA Model

- When analyzing data, the following equation reminds us that we look for an overall pattern and deviations from it:
  DATA=FIT + RESIDUAL.
- In the regression model, the FIT was the population regression line, and the RESIDUAL represented the deviations of the data from this line.
- The ANOVA model assued that the population standard deviations are all equal. ANOVA procedures are not extremely sensitive to unequal standard deviations, a formal check for equality of standard deviations is not recommended as a preliminary to the ANOVA. Instead, we use the following rule as a guideline.
- If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumptions of the equal standard deviations, and our results will still be approximately correct.

- When we assume that the population standard deviations are equal, each sample standard deviation is an estimate of σ.
- To combine these into a single estimate, we use a generalization of the pooling method.

$$s_p^2 = \frac{(n_1 - 1) \times s_1^2 + (n_2 - 1) \times s_2^2 + ... + (n_k - 1) \times s_k^2}{((n_1 - 1) + (n_2 - 1) + ... + (n_k - 1))} \tag{2}$$

  is an unbiased estimator of σ^2. The pooled standard deviation $s_p = \sqrt{s_p^2}$ is the estimate of σ.
- Pooling gives more weight to groups with larger sample sizes. If the sample sizes are equal s_p^2 is just the average of the k sample variances. Note that, s_p is not the average of the k sample standard deviations.

Testing hypothesis in one-way ANOVA

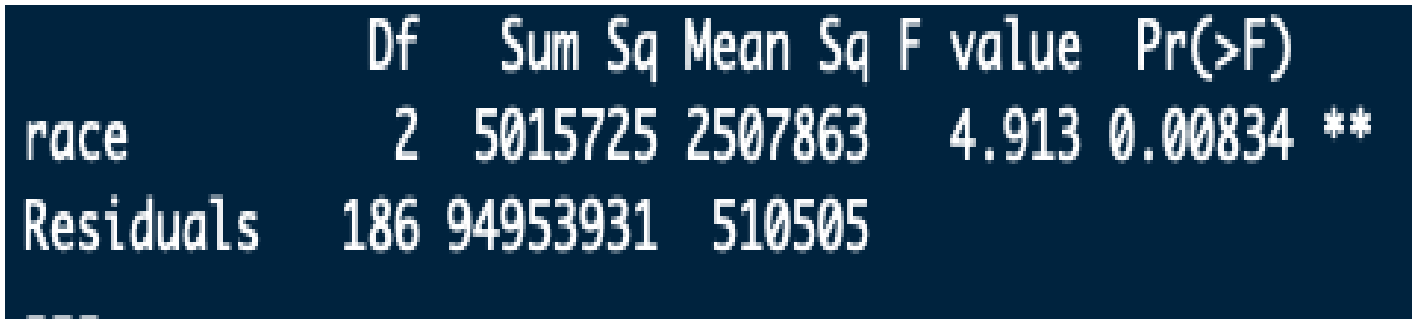
- Comparison of several means is accomplished by using an F statistic to compare the variation among groups with the variation within groups.
- The hypothesis for one-way ANOVA are:
  H_0 : $\mu_1 = \mu_2 = \ldots = \mu_n$
  H_a : not all of the μ_i are equal.

ANOVA

bwt

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5015725.25	2	2507862.63	4.913	.008
Within Groups	94953930.6	186	510505.003		
Total	99969655.8	188			

Figure 4: One-way ANOVA. SPSS output for understanding.



	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
race	2	5015725	2507863	4.913	0.00834	**
Residuals	186	94953931	510505			

Figure 5: One-way ANOVA. R output console.

1. SST = 99969656. SSG= 5015725. SSW=94953931. Note that:
   SST (Total sum of squares) = SSG (Sum of squares group) + SSW (sum of squares within)
2. The results of ANOVA indicate that there is sufficient evidence to reject the null hypothesis. $F(2,186) = 4.913$ with a p-value of 0.0083 ($p<0.05$).
3. This means that the data provide clear evidence to support the claim that these three groups of ethnicity have different mean birth weight values.

General view of ANOVA results can be represented as:

Table 2: Estimates of odds-ratio and relative risk in the given data set for objective poverty and low birth-weight.

	Df	Sum of squares	Mean Square	F Value
Group	df_G= k-1	SSG =$\sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$	MSG=$\frac{1}{df_G}$SSG	F=$\frac{MSG}{MSE}$
Residuals	df_E= n-k	SSE=$\sum_{i=1}^{k}(n_i - 1)s_i^2$	MSE=$\frac{1}{df_E}$ SSE	
Total	$df_T = n - 1$	SST=$\sum_{i=1}^{k}(x_i - \bar{x})^2$	MST=$\frac{1}{df_T}$ SST	

df = degree of freedom
SS=Sum of Squares
MS=Mean Square
F = ratio of variability in the sample means relative to variability within the groups.
k= number of groups.
n=total number of observations.
n_i = number of observations in group i
$\bar{x}_i$ = the mean of n_i observations in group i
$\bar{x}$ = the grand mean of all n observations.
s_i^2 = the variance of the n_i observations in group i.
The ANOVA table separates variation in the data into two parts: that Between Groups and the other Within Groups. In place of between groups, some software packages use Model or the name of the factor. Similarly, term like Error and Residual are frequently used in place of Within Groups.

- The Between Groups rown in the table corresponds to the FIT term in our DATA=FIT + RESIDUAL way of thinking. It gives information related to the variation **among** group means.
- The Within Groups row in the table corresponds to the RESIDUAL term in our DATA=FIT + RESIDUAL model. It gives information related to the variation **within** groups.
- Thus, DATA=FIT + RESIDUAL translates to Total=Between Groups + Within Groups.
- The fact is true in general. The total variation is always equal to the among-group variation plus the within-group variation.
- Associated with each sum of squares is a quantity called the degrees of freedom. Because SST measures the variation of all N observations around the overall mean, its degrees of freedom are DFT=N-1. This is the same as the degrees of freedom for the ordinary sample variance.
- Similarly because SSG measures the variation of the k sample means around the overall mean, its degrees of freedom are DFG=k-1.
- Finally, SSE is the sum of squares of the deviations $x_{ij} - \bar{x}_i$. Here we have N observations being compared with k sample means and DFE = N-k.

- Note that the degrees of freedom add in the same way that the sums of squares add. That is, DFT=DFG+ DFE.
- For each source of variation, the mean square is the sum of squares divided by the degrees of freedom.
- We can use the error mean square to find s_p, the pooled estimate of the parameter σ of our model. It is true in general that:
  $s_p^2 = MSE = \frac{SSE}{DFE}$.
  In other words, the error mean square is an estimate of the within-group variance, σ^2. The estimate of σ is therefore the square root of this quantity. So,
  $s_p = \sqrt{MSE}$.
  In our case, MSE $= 510505$, $\sqrt{MSE} = 714.4$

The F Test

- If H_0 is true, there are no differences among the group means. The ratio MSG.MSE is a statistic that is approximately 1 if H_0 is true and tends to be larger if H_a is true. This is the ANOVA F-statistic.
- In our example MSG = 2507863 and MSE = 510505, so the ANOVA F-statistic is $F = \frac{MSG}{MSE} = \frac{2507863}{510505} = 4.9125$
- When H_0 is true, the F statistic has an F distribution that depends upon two numbers: the degrees of freedom and the numerator and the degrees of the freedom for the denominator. These degrees of freedom are those associated with the mean squares in the numerator and the denominator of the F statistic.
- For one-away ANOVA, the degrees of freedom for the numerator are DFG=k-1, and the degrees of the freedom for the denominator are DFE=N-1. We use the notation F(k-1, N-1) for this distribution.

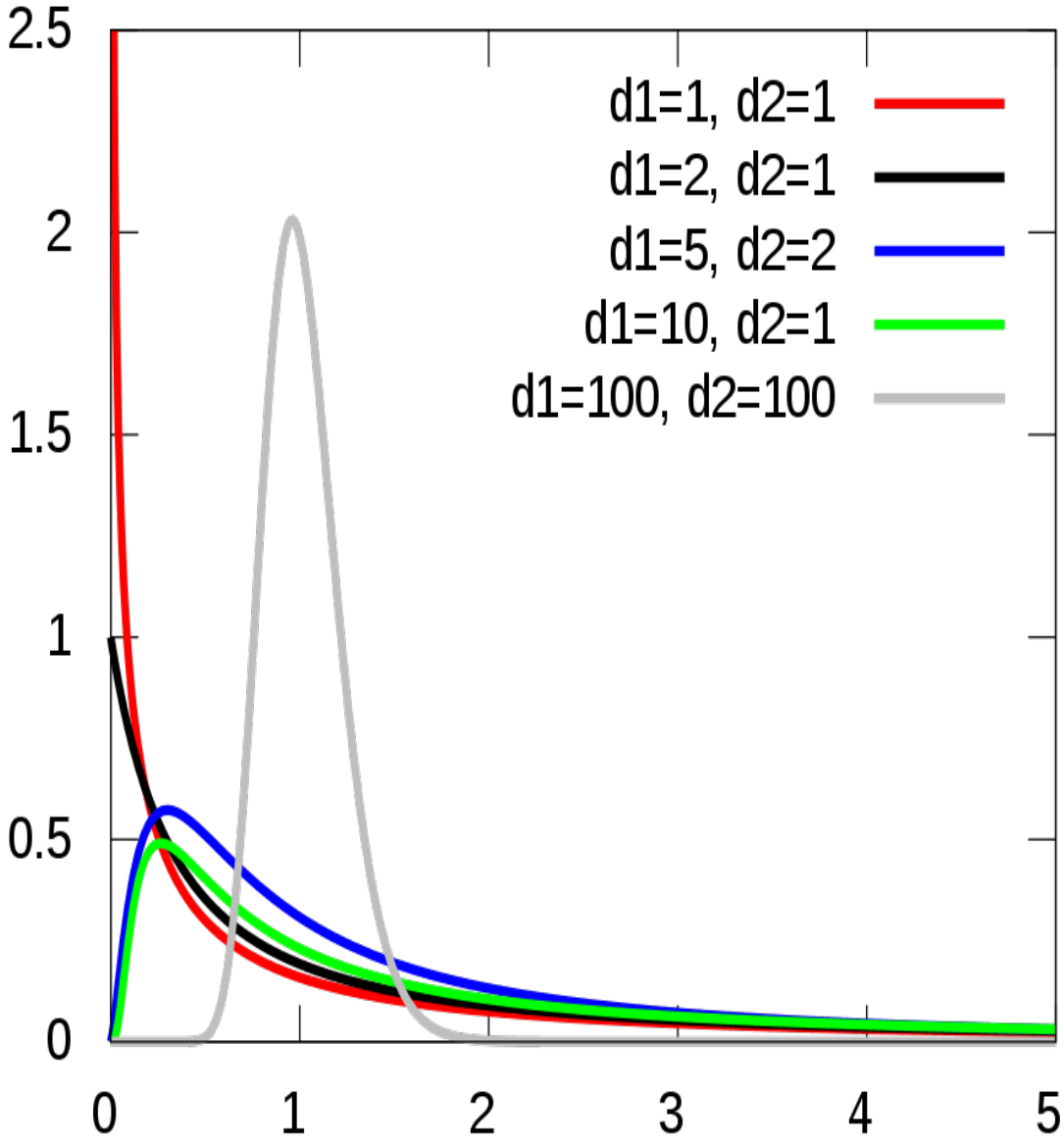


Figure 6: The probability distribution function of F distribution.

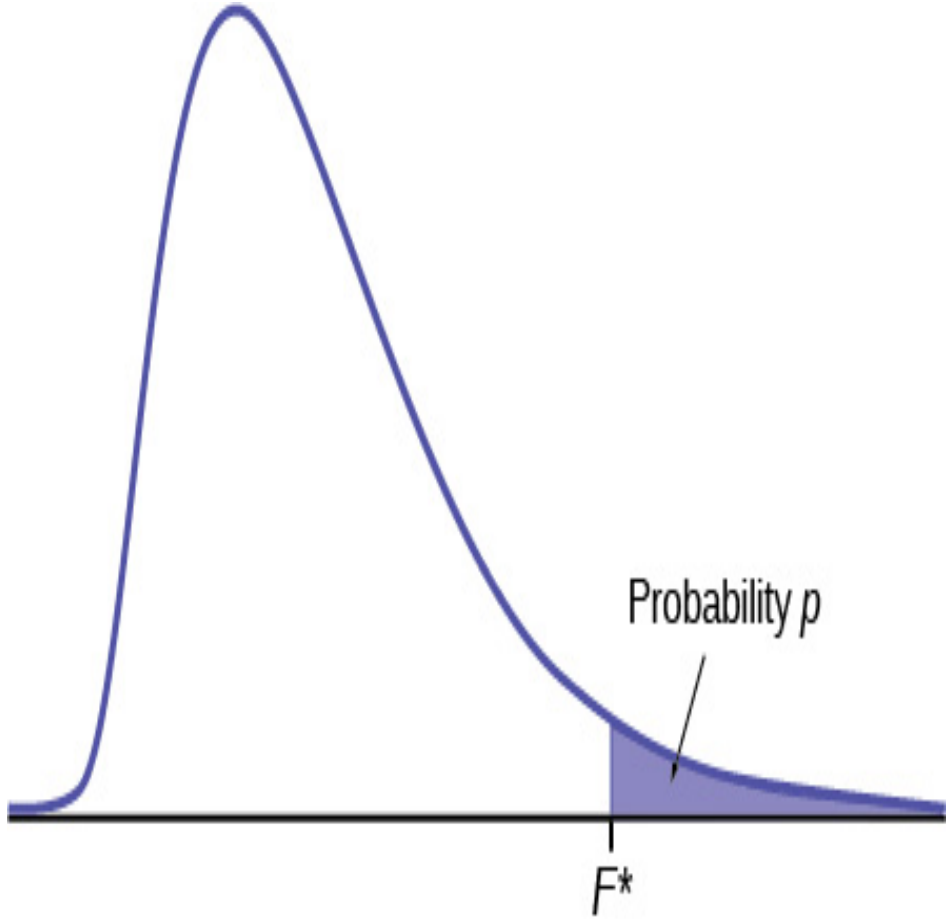


Figure 7: F distribution showing critical region.

- In the birth weight example, we found F=4.91. (Note that it is standard practice to round F statistics to two places after the decimal point). There were 3 populations, so the degrees of the freedom in the numerator are DFG=3-1=2.
- The degrees of freedom of denominator for this example are DFE=N-k=189-3=186.
- From table of F distribution (https://www.stat.purdue.edu/ jtroisi/STAT350Spring2015/tables/FTable.pdf) we first find the column corresponding to 2 degrees of freedom in the numerator.
- For the degrees of freedom in the denominator, we see that there are entries of 100 and 200. These values are very close. To be conservative we use critical values corresponding to 100 degrees of freedom in the denominator since these are slighly larger.

Table 3: Table listing the dummy variables used in the regression analysis.

p	Critical value
0.100	2.76
0.050	3.94
0.025	5.18
0.010	6.90
0.001	11.50

- We have F=4.91. This is very close to critical value for P=0.05. Using the table, however, we can conclude only that P<0.025. (Note that the more accurate calculations performed by software indicated that in fact, P=0.008 which is less than 0.025.)
- For this example, we reject H_0 and conclude that the population means are not all the same.
- Remember that the F test is always one-sided because any differences among the group means tend to make F large.
- The ANOVA F test shares the robustness of the two-sample t test. It is relatively insensitive to moderatenon-normality and unequal variances, especially when the sample sizes are similar.

- One other item given by some software for ANOVA is worth noting. For an analysis of variance, we define coefficient of determination as:
  $R^2 = \frac{SSG}{SST}$
  The coefficient of determination plays the same role as the squared multiple correlation R^2 in a multiple regression. We can easily calculate the value from ANOVA table entries.
  $R^2 = \frac{SSG}{SST} = \frac{5015725.25}{99969655.8} = 0.050$
- About 5% of variation in birth weight is explained by ethnicity of the mothers, White, Black and others. The other 95% of the variation is due to baby-to-baby variation within each of the three groups.
- We can see this in the histograms of Fig. 2. Each of the groups has a large amount of variation and there is a substantial amount of overlap in the distributions.
- The fact that we have evidence (P=0.008) against the null hypothesis that the three population means are not all the same does not tell us that the distribution of values are far apart.

Next in discussion Contrasts and Multiple comparisons.

Comparing the means: Contrasts

- The ANOVA F test gives a general answer to a general question: Are the differences among observed group means significant? Unfortunately, a small P-value simply tells us that the group means are not all the same.
- It does not tell us specifically which means differ from each other.
- Plotting and inspecting the means give us some indication of where the differences lie, but we would like to supplement inspection with formal inference.
- In the ideal situation, specific questions regarding comparisons among the means are posed before the data are collected.
- In the current example, $\bar{x_w} = 3102, \bar{x_b} = 2719, \bar{x_o} = 2805$.
- The null hypothesis tested was: $H_0 : \mu_w = \mu_b = \mu_o$ versus the alternative that the three population means are not all the same. We would report these results as F(2,186)=4.91, p=0.0083. Because the p-value is less than 0.05, we conclude that the data provide clear evidence that the three population means are not all the same.
- Having evidence that the three population means are not the same does not really tell us anything useful. We would really like our analysis to provide us with more specific information. The alternative hypothesis is true if:

  $\mu_w \neq \mu_b$
  or if $\mu_w \neq \mu_o$
  or if $\mu_b \neq \mu_o$

  or if any combination of these statements is true. *When you reject the ANOVA null hypothesis, additional analyses are required to obtain useful results.*

- Experts on birth weight and socioeconomic position would suggest that babies belonging to Black ethnicity face a very different environment than the White ethnicity babies. Therefore, a reasonable question to ask is whether or not the babies belonging to White ethnicity is different from the others.
- We can take this question and translate it into a testable hypothesis.
- To compare the White babies with the other two groups of babies we construct the following null hypothesis:

  $H_{01} : \frac{1}{2}(\mu_b + \mu_o) = \mu_w$

  We could use the two sided alternative

  $H_{a1} : \frac{1}{2}(\mu_b + \mu_o) \neq \mu_w$

  but we could also argue that the one-sided alternative

  $H_{a1} : \frac{1}{2}(\mu_b + \mu_o) < \mu_w$

  is appropriate for this problem because we expect the Black and other ethnic categories to have a household environment that is less facilitating than the White category households.

- In the example above, we used H_{01} and H_{a1} to designate the null and alternative hypotheses. We use H_{02} and H_{a2} for these hypotheses.
- H_{02} : $\mu_o = \mu_b$

  H_{a2}: $\mu_o \neq \mu_b$.
- Each of H_{01} and H_{02} says that a combination of population means is 0. These combinations of means are called **contrasts**.
- We use ψ, the Greek letter psi for contrasts among population means. For comparing the White babies with the other two groups of workers, we have

  $$\psi_1 = -\frac{1}{2}\ (\mu_b + \mu_o)\ + \mu_w$$

  $$= (-0.5)\mu_b + (-0.5)\mu_o + 1\mu_w$$

  and for comparing the Black babies with Other category babies

  $$\psi_2 = \mu_o - \mu_b$$
- In each case, the value of the contrast is 0 when H_0 is true.
- Note that the contrasts are chosen to be defined as positive so that they will be positive when the alternative interest (what we expect) is true.

- A contrast expresses an effect in the population as a combination of population means. To estimate the contrast, form the corresponding sample contrast by using sample means in place of population means.
- Under the ANOVA assumptions, a sample contrast is a linear combination of independent normal variables and therefore has a normal distribution.
- The standard error of a contrast can be obtained by using the rules of variances. Inference is based on t statistics.
- **Rules for Variances**

- **Rule 1:** if X is a random variable and a and b are fixed number, then

  $\sigma^2_{a+bX} = b^2\sigma^2_X$

- **Rule 2**: If X and Y are independent random variables, then

  $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$

  $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$

  This is the additional rule for variances of independent random variables.

- **Rule 3**: If X and Y have correlation ρ

  $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y + 2\rho\sigma_X\sigma_Y$

  $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y - 2\rho\sigma_X\sigma_Y$

Contrasts

- A contrast is a combination of population means of the form
  $\psi = \sum a_i \mu_i$
  where the coefficients a_i have sum 0. The corresponding sample contrast is
  $c = \sum a_i \bar{x}_i$
  The standard error of c is
  $SE_c = s_p \sqrt{\sum \frac{a_i^2}{n_i}}$
  To test the null hypothesis
  $H_0 : \psi = 0$
  use the t statistic
  t=$\frac{c}{SE_c}$
  with degrees of freedom DFE that are associated with s_p. The alternative hypothesis can be one-sided or two-sided.
  A level C confidence interval for ψ is
  $c \pm t * SE_c$
  where t* is the value for the t(DFE) density curve with area C between -t* and t*.

- In our example the coefficients for contrasts are $a_1 = -0.5$, $a_2 = -0.5$, $a_3 = 1$ for ψ_1 and $a_1 = 0, a_2 = 1, a_3 = -1$ for ψ_2, where the subscripts 1, 2, 3 correspond to μ_w, μ_o, μ_b respectively. In each case the sum of a_i is zero. Now let us look at the inference for each of these contrasts in turn.
- The sample contrast that estimates ψ_1 is
  $c_1 = (-0.5)\bar{x_b} + (-0.5)\bar{x_o} + 1\bar{x_w}$
  =(-0.5) 2719 + (-0.5)2805 + 3102 = 339
  with standard error
  $SE_{c1} = s_p\sqrt{\sum \frac{a_i^2}{n_i}}$
  $s_p^2 = \frac{(96-1)(727.89)^2+(26-1)(638)^2+(67-1)(722)^2}{(96-1)+(26-1)+(67-1)} = 510290.9$
  $s_p = \sqrt{510290.9} = 714.34$ Thus,
  $SE_{c1} = 714.34\sqrt{\frac{-0.5^2}{26} + \frac{-0.5^2}{67} + \frac{1^2}{96}} = 110.01$
  The t statistic for testing $H_{01} : \psi_1 = 0$ versus $H_{a1} : \psi_1 > 0$ is
  t=$\frac{c_1}{\psi_1}$
  $= \frac{339}{110.01} = 3.08$
  s_p has 186 degrees of freedom, software using the t(186) distribution gives the one-sided P-value as 0.001. The P-value is small, so there is strong evidence against H_0.

- We have evidence to conclude that the mean birth weight of White babies is higher than the average of the birth weights of babies in Black and other category.
- The size of the difference can be described with a confidence interval.
- $c_1 \pm t * SE_{c_1} = 339 \pm (1.984)(110.01)$
  $= 339 \pm 218.4$
  The interval is (120.6,557.4). We are 95% confident that the difference is between 120.6 gm and 557.4gm.

We use the same method for the second contrast

- $c_2 = \mu_o = \mu_b = 2805 - 2719 = 86$ with standard error $SE_{c_2} = 714.34\sqrt{\frac{1^2}{67} + \frac{-1^2}{26}} = 165$
- The statistic for assessing the significance of this contrast is t$=\frac{c_2}{SE_{c_2}} = \frac{86}{165} = 0.52$
- The P-value for the two-sided alternative is 0.604. The data do not provide us with evidence in favor of a difference in population mean birth weight between Black and Other groups.
- Let us look at the confidence interval. A confidence interval will tell us what values of the population difference are compatible with the data.
- $c2 \pm 1.984 * 165 = 86 \pm (1.984)(165) = 86 \pm 327.36 = (-241.36, 413.36)$. The interval is (-241.36, 413.36). With 95% we state that the difference between the population means for these two groups of babies is between -241.36 gms and 413.36 gms.
- Some statistical software packages report the test statistics associated with contrasts ad F statistics rather than t statistics. These F statistics are the squares of the t statistics described above.

Software output SPSS and R

Contrast Coefficients

Contrast	race 1	2	3
1	1	-.5	-.5
2	0	-1	1

Contrast Tests

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
bwt	Assume equal variances	1	340.23	110.142	3.089	186	.002
		2	85.59	165.089	.518	186	.605
	Does not assume equal variances	1	340.23	106.712	3.188	130.537	.002
		2	85.59	153.211	.559	51.190	.579

Figure 8: Output of planned contrasts SPSS

Software output SPSS and R

```
Call:
aov(formula = bwt ~ race, data = d)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.28  -502.72   -12.72   526.28  1887.28

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2875.90      60.16  47.805  < 2e-16 ***
racecontrast1 -226.82      73.43  -3.089  0.00232 **
racecontrast2   42.80      82.54   0.518  0.60476
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 714.5 on 186 degrees of freedom
Multiple R-squared:  0.05017,   Adjusted R-squared:  0.03996
F-statistic: 4.913 on 2 and 186 DF,  p-value: 0.008336
```

Figure 9: Output of planned contrasts R

Interpretation of the output

An ANOVA revealed that there was a significant association between ethnicities and birth weight of babies ($F(2,186) = 4.91$, $p=0.0083$). Further investigation of the ethnicities using planned contrasts revealed that a baby belonging to white ethnicity was associated with significant increase in birth weight compared to the black and other ethnicity ($t186)=-3.089$, $p=0.0023$ (one-tailed), and that there was no statistically significant difference found between birth weight of black and other ethnicity babies, $t(186)=0.52$, $p=0.604$ (two-tailed).
How to perform contrasts in SPSS and R

- For SPSS please refer to this **Contrast SPSS** link.
- For R please refer to this **Contrast R** link.

Multiple Comparisons

- To many studies, specific questions cannot be formulated in advance of the analysis. If H_0 is not rejected, we conclude that the population means are indistinguishable on the basis of the data given.
- On the other hand, if H_0 is rejected, we would like to know which pairs of means differ.
- **Multiple comparisons** methods address this issue. It is important keep in mind that multiple-comparison methods are used *only after rejecting the ANOVA* H_0.
- Return once more to the birth weight of babies example with three groups of ethnicities. We can make three comparisons between pairs of means: White versus Black, White versus Other and Black versus Other. We can write a t-statistic for each of these pairs,

$$t_{wb} = \frac{\bar{x_w} - \bar{x_b}}{s_p \sqrt{\frac{1}{n_w} + \frac{1}{n_b}}}$$

$$= \frac{3102 - 2719}{714.34 \sqrt{\frac{1}{96} + \frac{1}{26}}}$$

$$= 2.425$$

compares the populations of White and Black.

- The subscripts on t specify which groups are compared. The t-statistics for the other two pairs are:

$$t_{wo} = \frac{\bar{x_w} - \bar{x_o}}{s_p \sqrt{\frac{1}{n_w} + \frac{1}{n_o}}}$$

$$= \frac{3102 - 2805}{714.34 \sqrt{\frac{1}{96} + \frac{1}{67}}}$$

$$= 2.61$$

and

$$t_{bo} = \frac{\bar{x_b} - \bar{x_o}}{s_p \sqrt{\frac{1}{n_b} + \frac{1}{n_o}}}$$

$$= \frac{2719 - 2805}{714.34 \sqrt{\frac{1}{26} + \frac{1}{67}}}$$

$$= -0.52$$

- We performed the third calculation when we analysed the contrast ψ_2 in the previous section on contrasts, because that contrast was μ_o-μ_b . These statistics are very similar to the pooled two-sample t statistic for comparing two popuation means.
- The difference is that now we have more than two populations, so each statistic uses the pooled estimator s_p from all groups rather than the pooled estimator from just the two groups being compared.
- This additional information about the common σ increases the power of the tests. The degrees of freedom for all of these statistics are DFE $= 186$, those associated with s_p.
- Because we do not have any specific ordering of the means in mind as an alternative to equality, we must use a two-sided approach to the problem of deciding which pairs of mean are significantly different.

Mutliple Comparisons

- To perform a multiple-comparisons procedure, compute t statistics for all pairs of means using the formula

$$t_{ij} = \frac{\bar{x}_i - \bar{x}_j}{s_p \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}} \tag{3}$$

  If $\left|t_{ij}\right| \geq t^{**}$ we declare that the population means μ_i and μ_j are different. Otherwise, we conclude that the data do not distinguish between them. The value of t^{**} depends upon which multiple comparisons procedure we use.

- One obvious choice for $t{*}{*}$ is the upper $\alpha/2$ critical value for the t(DFE) distribution. This choice simply carries out as many separate significance tests of fixed level α as there are pairs of means to be compared.

- The procedure based on this choice is called the **least-significant differences method,** or simply LSD. It has some undesirable properties, particularly if the number of means being compared is large.

- For example, that there are k=20 groups and we use LSD with $\alpha = 0.05$. There are 190 different pairs of means. If we perform 190 t-tests, each of with an error rate of 5%, our overall error rate will be unacceptably large. We expect about 5% of the 190 to be significant even if the corresponding population means are the same. Since 5% of 190 is 9.5, we expect 9 or 10 false rejections.

- The LSD procedure fixes the probability of a false rejection for each single pair of means being compared. It does not control the overall probability of some false rejections among all pairs. Other choices of t^{**} control possible errors in other ways. The choice of t^{**} is a complex problem and there are different ways to adjust the t^{**}.
- One way s called the **Bonferroni method**.
- Use of this procedure with $\alpha = 0.05$ guarantees that the probability of any false rejection among all comparisons made is no greater than 0.05. This is much stronger protection than controlling the probability of a false rejection at 0.05 for each separate comparison.
- We apply Bonferroni multipl-comparison procedure with $\alpha = 0.05$ to the data with the t^{**} for this procedure is 2.14 (using software: R command : qt(c(0.05/3,1-0.05/3), df=186)).
- Of the statistics $t_{wb} = 2.425$, $t_{wo} = 2.61$ and $t_{bo} - 0.52 =$ calculated in the beginning of this section, only t_{wb} and t_{wo} are significant.
- Let us see what SPSS and R provide.

Multiple Comparisons

Dependent Variable: bwt

Bonferroni

(I) race	(J) race	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	383.026*	157.964	.049	1.41	764.64
	3	297.435*	113.742	.029	22.65	572.22
2	1	-383.026*	157.964	.049	-764.64	-1.41
	3	-85.591	165.089	1.000	-484.42	313.23
3	1	-297.435*	113.742	.029	-572.22	-22.65
	2	85.591	165.089	1.000	-313.23	484.42

*. The mean difference is significant at the 0.05 level.

Figure 10: Output generated by SPSS of multiple comparison tests using Bonferroni method.

- The output generated by SPSS for Bonferroni comparisons appears in Fig. 10. The software uses an asterisk to indicate that the difference in a pair of means is statistically significant. These results agree with the calculations we performed in the example above using modified t value.
- Note that each comparison is given twice in the output.
- Note that the confidence intervals of the difference is also provided for the multiple comparisons.

Tukey Method

- One more method used generally as a post-hoc comparison is the Tukey Honest Significant Differences (HSD). Although it is generally not a post-hoc analysis method because it can be used without using ANOVA also.
- The test is known by several different names. Tukey's test compares the means of all treatments to the mean of every other treatment and is considered the best available method in cases when confidence intervals are desired or if sample sizes are unequal.
- For more information on Tukey test, please follow this link Post-Hoc Analysis with Tukey's Test

Multiple Comparisons

Dependent Variable: bwt

Tukey HSD

(I) race	(J) race	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	383.026*	157.964	.043	9.82	756.24
	3	297.435*	113.742	.026	28.71	566.17
2	1	-383.026*	157.964	.043	-756.24	-9.82
	3	-85.591	165.089	.862	-475.63	304.45
3	1	-297.435*	113.742	.026	-566.17	-28.71
	2	85.591	165.089	.862	-304.45	475.63

*. The mean difference is significant at the 0.05 level.

Figure 11: Output generated by SPSS of multiple comparison tests using Tukey method.

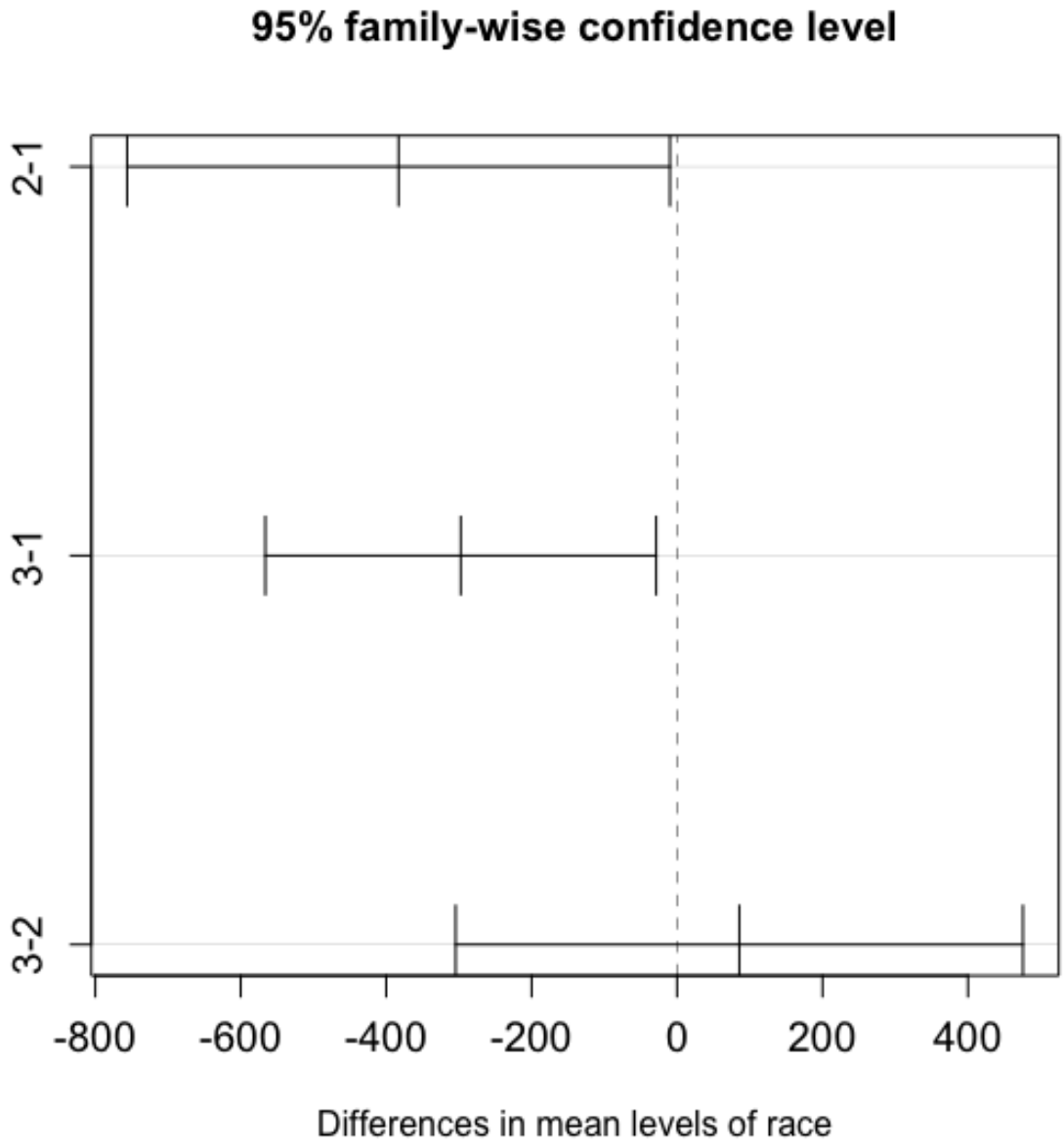


Figure 12: Pairwise confidence interval given by Tukey. 1 indicates White, 2 indicates Black and 3 indicates Other. (Software used: R.)

Inputs takes majorly from Introduction to the Practice of Statistics, Fifth edition. David S. Moore. George P. McCabe.